OR2009 Presentation Proposal: Author Identifiers in Scholarly Repositories

Simeon Warner
Cornell Information Science and
Cornell University Library
Ithaca, NY 14850, USA
simeon@cs.cornell.edu

Bibliometric and usage-based analyses and tools highlight the value of information about scholarship contained within the network of authors, articles and usage data. Less progress has been made on populating and using the author side of this network than the article side, in part because of the difficulty of unambiguously identifying authors. I briefly review a sample of author identifier schemes, and consider use in scholarly repositories. I then describe work at arXiv to implement public author identifiers, services based on them both locally and through a Facebook application, and plans to make this information useful beyond the boundaries of arXiv.

1 Context

In an ideal scholarly communication system there would be tools to browse, navigate, make recommendations and assess influence based on the complete graph of all actors (people, collaborations, institutions) and all communication artifacts (articles, comments, blog posts, usage data[1]). As a shorthand I will call this complete graph the *publication network*. Contained within it are the familiar citation, usage, co-authorship, and co-citation graphs. In recent bibliometric and usage-based work, significant progress has been made with the artifact part of this graph (see, for example the work of the MESUR project [3]). Much less progress has been made with the actor part of the graph, in part because it is much harder to unambiguously identify authors than articles.

In a recent Nature Correspondence, Raf Aerts asked *"If it is possible to have DOIs for objects (or, so they say, enough IPv6 addresses for every molecule on Earth), why is it so difficult to implement DAIs for authors?"* [1]. Raf had earlier hinted at part of the answer by pointing out that he has more than one identifier in Scopus [5]. That is, it is difficult to mine existing data to disambiguate references to authors. The more fundamental answer is that it is much easier to create DOIs for articles when the one owner for the article creates the one DOI for it and presents it with the article (ignoring the issue of multiple versions of articles). As authors, we are not owned by a single authority and even if an identifier were created for us at birth by the appropriate government, there would be significant privacy concerns about using it for everything (cf. social security numbers in the USA). While we want to link a single author's works together, do we want that identity to immediately link us to all other digital information about the private life of the individual?

[1] Logically usage data would be links between actors and artifacts. However, for historical, cultural and practical reasons most usage data is treated as anonymous even if co-usage information is extracted.

2 Author Identifiers

The arguments above and the understanding that there are many different interests in, and uses for, author/person identities, suggest that there will many different systems and multiple identities for each author. In the scholarly communication domain there will be a patchwork of overlapping publication networks. Combining these networks will be vastly easier than assembling the combined network from raw textual data. A significant aid will be the addition of assertions that link identities in different networks (e.g. Author1 in Network2 is the same person as AuthorX in NetworkY), expressed either via the Semantic Web or repository metadata.

Table 1 shows some example schemes used for scholarly author identifiers. OpenID is not limited to the scholarly domain and is aimed primarily at authentication. However, if it continues to see growing acceptance it may well be a useful open system that repositories could use. The largest efforts to create author identifiers specifically for the scholarly domain, Scopus Author Identifiers and ResearcherID, come from commercial entities and are clearly motivated by the desire to provide improved services based upon then. It is not clear how open the interfaces based on these identifiers will be, or what data about them will be openly available. However, even if relatively little is shared, understanding the mappings to identifiers in these systems may provide common join points in other repositories that do openly share data.

Scheme	Example	Authority
OpenID	`http://samruby.myopenid.com/`	anyone supporting the protocol
Scopus Author Id	7103063073	Elsevier
ResearcherID	A-1637-2009	Thomson Reuters
Digital Author Id	info:eu-repo/dai/nl/304825271	Dutch Universities and Research Institutes
arXiv Author Id	`http://arxiv.org/a/warner_s_1`	arXiv.org

Table 1: A sample of identifier schemes used for scholarly author identifiers

The two other examples in table 1 illustrate decreasing scopes. In the Netherlands the Digital Author Id (DAI) *"is a unique national number assigned to every author who has been appointed to a position at a Dutch university or research institute or has some other relevant connection with one of these organizations"* [6]. The DAI provides a join point for data in different repositories and enables services based on this combined data (e.g. NARCIS). The arXiv author identifier is local to a single repository but the ability to record foreign identifiers in the author record (say a DAI), and the ability to match articles between repositories, will allow the joining of data across repository boundaries.

3 Author Identifiers at arXiv

There are a significant number of physicists for whom all articles or at least all recent articles are available on arXiv. It is not uncommon to find web homepages with a link to arXiv author search in place of a bibliography — why maintain the information in a second place when arXiv will do it automatically? Fielded author search has been used in this way for many years and has exactly the same problems of author disambiguation as text-based efforts to build the publication network.

With the introduction of user accounts, arXiv, like many other repositories, started to collect data on which user made each submission and whether he or she claimed to be an author. This start to building authority records was augmented by attempts to retrospectively associate older papers with users based on email address matching, and the introduction of facilities by which users could "claim ownership" of existing submissions. Various heuristics are used to limit what papers can be claimed automatically. Use was motivated through the introduction of an endorsement system[2] where users must be known as authors of a certain number of papers in order to endorse new users.

Demanding identification of all authors at submission time was considered impractical. For articles with one or two authors identification would not be too burdensome, but for papers with 10 or even 2500 authors[3] it is clearly impractical. A solution that uses arXiv administrator effort to deal with each article is also impractical because just two administrators handle all user queries relating to arXiv's 55,000 submissions/year — most submissions must be entirely automated. We thus decided on an approach that will create useful services based on a public author identifier which we internally link to our user records. We hope that by providing useful services our users will be motivated to further improve the authority records on which these services depend.

3.1 Author URI and Services

We have opted for a web-centric approach using Linked Data [2] style access. Each arXiv author identifier is a URI (e.g. `http://arxiv.org/a/warner_s_1`) which supports HTTP content-negotiation. By default, or if selected via content-negotiation headers, the arXiv author URI redirects to an HTML page listing all arXiv publications authored by the given individual based on our user records. This already solves the problem of name collision in arXiv author search and so provides a more reliable link than our text-based author search. Such a list on the arXiv site would still be an extra click away from the user's homepage. We thus provide JavaScript that a user may include in their homepage to dynamically include an up-to-date publication list from arXiv. Various formatting options are provided and the content may be styled using CSS. This facility is based upon a content-negotiated request for an Atom representation of the arXiv author id resource which results in a machine readable Atom feed of paper information (in the same format as the arXiv API[4]).

arXiv's second use of arXiv author ids is to leverage this automatically generated and updated list of publications to lower the effort required to integrate arXiv papers into social networking sites. Facebook was chosen as the first site to work with but the OpenSocial API is also being investigated. Once the arXiv Facebook application has been told the association between a user's Facebook account and their arXiv author identifier, a list of publications is immediately available as either a panel or a tab on their Facebook profile — all title, author list, abstract and linking information is automatically imported from arXiv. New or old publications may be reported in the user's feed, with optional comments, and thus show up in friends' news feeds. This application is being tested and is scheduled for release in mid-March 2009. We cannot predict how many people will use the application but will carefully study and report uptake.

[2] `http://arxiv.org/help/endorsement`

[3] Articles from high-energy physics collaborations often have many authors. See, for example, the recent ATLAS collaboration paper `http://arxiv.org/abs/0901.0512` with > 2500 authors

[4] `http://arxiv.org/api`

3.2 Helping to Build the Publication Network

arXiv is making the multiple-identifier problem one identifier worse by creating arXiv specific identifiers. It has been noted by several authors that deduping articles is a key problem in bibliometrics and we don't want to create the same situation for authors. However, because we are already supplying machine readable information it is possible to augment this with information with associations between author ids in different schemes. OpenID explicitly plans for multiple identifiers for a single person, or even for multiple ids for each chosen persona of a person, and provided facilities to express and leverage them in the Yadis/XRDS document [7]. We plan to use OAI-ORE Resource Maps [4] as an additional expression of the author record and to include information about other identities that have been registered by arXiv users.

4 Acknowledgements

I am pleased to acknowledge contributions from Nathan Woody (Facebook and JavaScript interface for arXiv), Thorsten Schwander and Paul Ginsparg. This work is supported by Microsoft through a Technical Computing Initiative (TCI) Grant.

References

[1] Raf Aerts. Digital identifiers work for articles, so why not for authors? *Nature*, 453, 2008. http://dx.doi.org/10.1038/453979b.

[2] Chris Bizer, Richard Cyganiak, and Tom Heath. How to Publish Linked Data on the Web, 2007. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/20070727/.

[3] Johan Bollen, Herbert Van de Sompel, and Marko A. Rodriguez. Towards Usage-based Impact Metrics: First Results from the MESUR Project. *Proceedings of the Joint Conference on Digital Libraries 2008 (JCDL08), June 16, 2008, Pittsburgh, Pennsylvania, USA*, 2008. http://arxiv.org/abs/0804.3791.

[4] Open Archives Initiative – Object Reuse and Exchange, 2007. http://www.openarchives.org/ore/toc.

[5] Scopus Author Identifier. http://www.info.scopus.com/. The main Scopus service http://www.scopus.com/scopus/home.url and the details of author identifier are accessible from http://tinyurl.com/yta9m7.

[6] SURFfoundation. Digital Author Identifier (DAI). http://www.surffoundation.nl/smartsite.dws?ch=eng&id=13480.

[7] Yadis 1.0: The Identity and Accountability Foundation for Web 2.0, 2006. http://yadis.org/wiki/Yadis_Documents.